J Optim Theory Appl (2016) 170:900–915
DOI 10.1007/s10957-016-0963-5

A Weighted Mirror Descent Algorithm for Nonsmooth Convex Optimization Problem

Duy V. N. Luong[1] · **Panos Parpas**[1] · **Daniel Rueckert**[1] · **Berç Rustem**[1]

Received: 23 March 2015 / Accepted: 31 May 2016 / Published online: 14 June 2016

Abstract Large-scale nonsmooth convex optimization is a common problem for a range of computational areas including machine learning and computer vision. Problems in these areas contain special domain structures and characteristics. Special treatment of such problem domains, exploiting their structures, can significantly reduce the computational burden. In this paper, we consider a Mirror Descent method with a special choice of distance function for solving nonsmooth optimization problems over a Cartesian product of convex sets. We propose to use a nonlinear weighted distance in the projection step. The convergence analysis identifies optimal weighting parameters that, eventually, lead to the optimally weighted step-size strategy for every projection on a corresponding convex set. We show that the optimality bound of the Mirror Descent algorithm using the weighted distance is either an improvement to, or in the worst case as good as, the optimality bound of the Mirror Descent using unweighted distances. We demonstrate the efficiency of the algorithm by solving the

Communicated by Amir Beck.

Electronic supplementary material The online version of this article (doi:10.1007/s10957-016-0963-5) contains supplementary material, which is available to authorized users.

✉ Duy V. N. Luong
vu.luong05@alumni.imperial.ac.uk

Panos Parpas
panos.parpas@imperial.ac.uk

Daniel Rueckert
d.rueckert@imperial.ac.uk

Berç Rustem
b.rustem@imperial.ac.uk

[1] Imperial College London, London, UK

Markov Random Fields optimization problem. In order to exploit the domain of the problem, we use a weighted log-entropy distance and a weighted Euclidean distance. Promising experimental results demonstrate the effectiveness of the proposed method.

Keywords Subgradient projection · Weighted distance · Mirror Descent · Markov Random Fields

Mathematics Subject Classification 90C06 · 90C25 · 90C35

1 Introduction

It is well known that convex optimization problems can be solved in polynomial time at a low iteration count using interior point methods. However, most of these methods do not scale well with the dimension of the optimization problem. A single iteration cost of an interior point method grows nonlinearly with the problem size. As a result, low iteration count becomes expensive in terms of computational performance. Since what matters most in practice is the overall computational time to solve the problem, first-order methods with computationally low-cost iterations become a viable choice for large-scale optimization problems. In this paper, we present an efficient first-order method to solve a general large-scale nonsmooth optimization problem over a Cartesian product of convex sets. The proposed method is the Mirror Descent (MD) algorithm [1–4], an iterative first-order approach for nonsmooth optimization problems, with a special choice of distance function. The main idea of MD is to utilize a suitable Bregman distance [5] and identify the optimal step-size for the projection step over the feasible domain. In the case where the domain is a Cartesian product of convex sets, we propose to use optimal step-size strategy for each projection on the corresponding subset instead of using a common step-size for the projection on the entire domain. In order to achieve this, we employ a weighted distance function for the projection scheme. The weighted distance function exploits the 'disjoint' property of the problem's domain by considering suitable *weights* for every subset. By assessing the optimality bound for the proposed algorithm, we establish the optimal weighting parameters for each distance function of the corresponding subset. These weighting parameters influence the projection step as scaling factors of the common step-size. Thus, the step-size is scaled appropriately for the corresponding subset projection.

As an illustration, we demonstrate the performance of the proposed algorithm, hereafter referred to as the weighted MD, by solving the Markov Random Fields (MRF) optimization problem [6,7]. This problem often arises from the areas of image analysis and machine learning [8]. We employ the weighted MD with log-entropy distances and optimal subset-dependent step-sizes to initialize the starting point. Subsequently, we use the weighted MD with Euclidean distances and incorporate the duality gap in the step-size computation. Experimental results confirm the superiority of the weighted MD over the MD algorithm with unweighted distance.

The remainder of this paper focuses on analyzing and describing the proposed weighted MD algorithm and its application to the MRF optimization problem. In the next section, we review the MD algorithm with a general distance function. Section 3

derives the optimality bound for solving a nonsmooth convex optimization problem over a Cartesian product of convex sets using MD. In addition, Sect. 3 introduces required definitions for developing the weighted MD algorithm. In Sect. 3.3, we derive the optimality bound of the proposed weighted MD algorithm and show that it is either an improvement to, or in the worst case as good as, the MD algorithm as described in Sect. 2. In Sect. 4, we consider the dual of the MRF optimization problem. The MRF dual belongs to the class of large-scale nonsmooth optimization problem over a Cartesian product of convex sets. We can therefore employ the weighted MD to solve it. We report very promising computational results in the online supplementary material provided.

2 Mirror Descent Algorithm

Consider the following nonsmooth convex optimization problem:

$$\max_{x\in\mathcal{X}} f(x), \tag{1}$$

where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$ is the Cartesian product of N closed and convex sets; and $\mathcal{X} \subset \mathbb{R}^n$. In this problem, the decision variable x can be decomposed into N disjoint blocks, where each block $x_i \in \mathcal{X}_i$. In addition, we assume the following for (1):

- The objective function $f : \mathcal{X} \to \mathbb{R}$ is concave and Lipschitz continuous.
- $f^* := f(x^*)$ denotes optimal objective value, where $x^* \in \mathcal{X}$.

Problem (1) can be solved by the Mirror Descent algorithm. MD algorithm [1–4] is a generalization of the projected subgradient method. The standard subgradient approach employs the Euclidean distance function with a suitable step-size in the projection step. Mirror Descent extends the standard projected subgradient method by employing a nonlinear distance function with an optimal step-size in the nonlinear projection step. In this section, we review the Mirror Descent algorithm for solving problem (1) without considering the domain geometry.

Let D(., .) denote the distance between any two points in the set $\mathcal{X}$, and MD algorithm employs a sequence of nonlinear projection:

$$x^{k+1} = \operatorname*{argmax}_{x\in\mathcal{X}} \left\langle f'_{x^k}, x \right\rangle - \frac{1}{\mu} \mathrm{D}(x, x^k), \tag{2}$$

where f'_{x^k} is a subgradient at the point x^k, μ is the optimal step-size. The set up of Mirror Descent requires D(., .) *compatible* with the norm:

- $\|.\|$ on the space embedding $\mathcal{X}$ and its dual norm:
- $\|\xi\|_* = \max_{x\in\mathcal{X}}\{\langle x, \xi \rangle : \|x\| \le 1\}$.

The maximum distance is given by $\Omega = \max_{x,y\in\mathcal{X}} \mathrm{D}(x, y)$. Suppose $f(x)$ is Lipschitz continuous on $\mathcal{X}$ with the Lipschitz constant $\mathcal{L} = \max_{x\in\mathcal{X}} \|f'_x\|_* < \infty$, we have the following convergence property for MD algorithm.

Theorem 2.1 *Let* f^* *denotes the global optimal objective function and* $\bar{x} = \text{argmax}_{x=\{x^1,\ldots,x^K\}} f(x)$. *Then, using the optimal step-size:*

$$\mu = \frac{\sqrt{2\,\Omega}}{\mathcal{L}\sqrt{K}}, \tag{3}$$

we have the following optimality bound after K *iterations:*

$$f^* - f(\bar{x}) \leq \frac{\mathcal{L}\sqrt{2\,\Omega}}{\sqrt{K}}. \tag{4}$$

Theorem 2.1 is a well-known result, and its proof can be found in [2,4]. In the following section, we derive explicitly the optimality bound where the domain $\mathcal{X}$ is the Cartesian product of subsets $\mathcal{X}_i, i = 1, 2, \ldots, N$. After that, we introduce a new distance function that will improve the derived optimality bound. The proposed parameterised distance naturally assigns weighting parameters to the projection step (2) on each subset $\mathcal{X}_i$.

3 Mirror Descent Algorithm with Weighted Distance

We consider a distance measurement on the given domain (the Cartesian product of many subsets) as a sum of weighted subset distances. In this setting, each subset is equipped with a specific distance function and a weighting parameter. We subsequently utilize this weighted distance in the projection step to develop a weighted Mirror Descent algorithm.

3.1 Weighted Distance Function

The distance function $\mathrm{D}(x, y)$ is defined as the Bregman distance:

$$\mathrm{D}(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle,$$

where $\psi(.)$ is σ-strongly convex over a *compatible* norm $\|.\|$, i.e.,

$$\langle \nabla\psi(x) - \nabla\psi(y), x - y\rangle \geq \sigma\|x - y\|^2, \quad \forall x, y \in \mathcal{X}. \tag{5}$$

Without any loss of generality, we assume[1] $\sigma = 1$ throughout the paper. A compatible norm $\|.\|$ is dependent of the choice of distance function. For example, l_1-norm is chosen for log-entropy distance [4], l_2-norm for Euclidean distance. Instead of using one distance function over the entire domain, let us consider separate choices of Bregman distance D_i for each subset $\mathcal{X}_i, i \in \{1, 2, \ldots, N\}$:

$$\mathrm{D}_i(x_i, y_i) = \psi^i(x_i) - \psi^i(y_i) - \langle \nabla\psi^i(y_i), x_i - y_i \rangle, \forall x_i, y_i \in \mathcal{X}_i. \tag{6}$$

[1] Note that Theorem 2.1 assumes $\sigma = 1$.

Each subset distance $\mathrm{D_i}(x_i, y_i)$ is equipped with a compatible norm $\|.\|_i$. Various choices of distance functions and compatible norms are discussed in [5,9,10]. Two examples that are relevant to the MRF application we consider later are:

- Euclidean distance: $\mathrm{D_i}(x_i, y_i) = \frac{1}{2}\|x_i - y_i\|_2^2$. In this case, $\psi^i(x_i) = \frac{1}{2}\|x_i\|_2^2$ and it is straightforward to show $\psi^i(.)$ is 1-strongly convex w.r.t $\|.\|_2$.
- Log-entropy distance: $\mathrm{D_i}(x_i, y_i) = \sum_j x_i^j \log(x_i^j / y_i^j) + y_i^j - x_i^j$. In this case, $\psi^i(x_i) = \sum_j x_i^j \log x_i^j - x_i^j$ is shown to be 1-strongly convex w.r.t. $\|.\|_1$ [4].

When $x, y \in \mathcal{X}$ and the domain $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$, the distance between x and y is equivalent to the sum of distances $\mathrm{D_i}(x_i, y_i)$. Using this definition, we can now state a corollary to Theorem 2.1.

Corollary 3.1 *Let Ω_i denote the maximum distance of a subset $\mathcal{X}_i$, i.e., $\Omega_\mathrm{i} = \max_{x_i, y_i \in \mathcal{X}_i} \mathrm{D_i}(x_i, y_i)$, and let $\mathcal{L}_\mathrm{i} = \max_{x_i \in \mathcal{X}_\mathrm{i}} \|f'_{x_i}\|_*$ denotes the local Lipschitz constant w.r.t. to a subset $\mathcal{X}_\mathrm{i}$. The optimality bound (4) for solving problem (1) by the Mirror Descent algorithm is given by:*

$$f^* - f(\bar{x}) \leq \frac{\sqrt{\sum_{i=1}^N \mathcal{L}_\mathrm{i}^2}\sqrt{2\sum_{i=1}^N \Omega_\mathrm{i}}}{\sqrt{K}}. \tag{7}$$

Proof When $\mathcal{X}$ is the Cartesian product of N convex sets $\mathcal{X}_i, i \in \{1, 2, \ldots, N\}$, the distance between two vectors $x, y \in \mathcal{X}$ is the sum of distances between any two blocks $x_i, y_i \in \mathcal{X}_\mathrm{i}$. As a result, the maximum distance Ω is also the sum of maximum distances on subset $\mathcal{X}_i$:

$$\Omega = \sum_{i=1}^N \Omega_\mathrm{i}\,. \tag{8}$$

Since the subsets $\mathcal{X}_i$ and $\mathcal{X}_j$ are independent, $i \neq j; i, j \in \{1, 2, \ldots, N\}$, we have:

$$\mathcal{L} = \max_{x \in \mathcal{X}} \|f'_x\|_* = \max_{x \in \mathcal{X}} \sqrt{\sum_{i=1}^N \|f'_{x_i}\|_*^2} = \sqrt{\sum_{i=1}^N \max_{x_i \in \mathcal{X}_i} \|f'_{x_i}\|_*^2} = \sqrt{\sum_{i=1}^N \mathcal{L}_\mathrm{i}^2}. \tag{9}$$

Substituting Ω and $\mathcal{L}$ in the optimality bound (4) yields (7). □

We now propose a weighted distance function in order to improve the optimality bound (7). For each subset distance $\mathrm{D_i}$, let us introduce a weighting parameter $\alpha_\mathrm{i} > 0$. The new distance function is then defined as a weighted combination of subset distances:

$$\mathrm{D}(x, y) := \sum_{i=1}^N \alpha_\mathrm{i}\, \mathrm{D_i}(x_i, y_i) = \sum_{i=1}^N \alpha_\mathrm{i}\, \psi^i(x_i) - \alpha_\mathrm{i}\, \psi^i(y_i) - \alpha_\mathrm{i} \langle \nabla \psi^i(y_i), x_i - y_i \rangle\,. \tag{10}$$

This yields the definition for $\psi(x)$ as a weighted sum of convex function $\psi^i(x_i)$:

$$\psi(x) = \sum_{i=1}^N \alpha_\mathrm{i}\, \psi^i(x_i). \tag{11}$$

Substituting (10) in the projection step (2) naturally yields:

$$x^{k+1} = \underset{x\in\mathcal{X}}{\operatorname{argmax}} \left\langle f'_{x^k}, x \right\rangle - \frac{1}{\mu} \sum_{i=1}^{N} \alpha_{\mathrm{i}}\, \mathrm{D}_{\mathrm{i}}(x_i, x_i^k). \tag{12}$$

Essentially, the property of $\mathcal{X}$ triggers an ability to independently compute the projection (12) on each subset $\mathcal{X}_{\mathrm{i}}$. In other words, if we consider the optimality condition of the optimization problem (12) w.r.t. each block $x_i \in \mathcal{X}_{\mathrm{i}}$, then (12) is separable and is equivalent to:

$$\forall i \in \{1, \ldots, N\}: \quad x_i^{k+1} = \underset{x_i\in\mathcal{X}_{\mathrm{i}}}{\operatorname{argmax}} \left\langle f'_{x_i^k}, x_i \right\rangle - \frac{\alpha_{\mathrm{i}}}{\mu}\, \mathrm{D}_{\mathrm{i}}(x_i, y_i). \tag{13}$$

As a result, we hope to achieve better performance by using suitable (or optimal) weighting parameters α_{i} for the corresponding subset $\mathcal{X}_{\mathrm{i}}$.

3.2 Compatible Norm, Dual Norm, Weighted Lipschitz Constant and Maximum Weighted Distance

In order to analyze the convergence of the sequence generated by (12), we need to establish the Lipschitz constant. This can be computed as the upper bound of the dual norm of the subgradients. To this end, we propose a *compatible* norm $\|.\|$ associated with the weighted distance.

Lemma 3.1 *For all $i \in \{1, \ldots, N\}$, let $\alpha_{\mathrm{i}} > 0$, $\psi^i(x_i)$ is 1-strongly convex w.r.t. $\|x_i\|_i$, and then, the weighted function, $\psi(x) = \sum_{i=1}^{N} \alpha_{\mathrm{i}}\, \psi^i(x_i)$, is 1-strongly convex w.r.t. the weighted norm:*

$$\|x\| := \sqrt{\sum_{i=1}^{N} \alpha^i \|x_i\|_i^2}. \tag{14}$$

Proof We have, $\forall x, y \in \mathcal{X}$:

$$\langle \nabla \psi(x) - \nabla \psi(y), x - y \rangle \geq \sum_{i=1}^{N} \alpha^i \|x_i - y_i\|_i^2 = \|x - y\|^2.$$

□

The dual norm $\|.\|_*$ of the proposed weighted norm (14) can be derived using the definition of dual norm (see Sect. 2 and [11]):

$$\|\xi\|_* = \sqrt{\sum_{i=1}^{N} \frac{\|\xi_i\|_{i*}^2}{\alpha_{\mathrm{i}}}}, \tag{15}$$

where $\|.\|_{i*}$ is a dual norm of $\|.\|_i$ over the subset $\mathcal{X}_\mathrm{i}$. Let $\mathcal{L}_\mathrm{i} = \max_{x_i \in \mathcal{X}_\mathrm{i}} \|f'_{x_i}\|_{i*}$ denote the local Lipschitz constant w.r.t. to a subset $\mathcal{X}_\mathrm{i}$; then, the weighed Lipschitz constant is given by:

$$\mathcal{L} = \max_{x \in \mathcal{X}} \|f'_x\|_* = \sqrt{\sum_{i=1}^{N} \frac{{\mathcal{L}_\mathrm{i}}^2}{\alpha_\mathrm{i}}}. \tag{16}$$

In addition, the maximum weighted distance Ω becomes:

$$\Omega = \max_{x,y \in \mathcal{X}} \mathrm{D}(x, y) = \sum_{i=1}^{N} \alpha_\mathrm{i}\, \Omega_\mathrm{i}, \tag{17}$$

where $\Omega_\mathrm{i} = \max_{x_i, y_i \in \mathcal{X}_i} \mathrm{D_i}(x_i, y_i)$.

Remark 3.1 The unweighted functions (8) and (9) in Sect. 2 can be viewed as a special case of the above-weighted functions where $\alpha_\mathrm{i} = 1\ ,\ \forall i = 1, 2, \ldots, N$.

3.3 Convergence Properties

We show the first result for optimality bound of the weighted MD algorithm.

Lemma 3.2 *Let f^* denote the global optimal objective function and $\bar{x} = \mathrm{argmax}_{x=\{x^1,\ldots,x^K\}} f(x)$ and μ be the step-size. We have the following optimality bound after K iterations:*

$$f^* - f(\bar{x}) \leq \frac{\Omega}{K\mu} + \frac{\mu\, \mathcal{L}^2}{2}. \tag{18}$$

Similar results can be found in [1,2,4]. The initial bound (18) depends on three terms μ, $\mathcal{L}$ and Ω, where the last two terms are themselves functions of the weighting parameters α_i. Therefore, we can tighten the bound (18) by considering its minimization w.r.t. μ and α_i.

Theorem 3.1 *For each subset $\mathcal{X}_\mathrm{i}$, let $\mathcal{L}_\mathrm{i} = \max_{x_i \in \mathcal{X}_\mathrm{i}} \|f'_{x_i}\|_{i*}$ be the local Lipschitz constant and $\Omega_\mathrm{i} = \max_{x_i, y_i \in \mathcal{X}_i} \mathrm{D_i}(x_i, y_i)$ be the maximum subset distance. Then, the optimal weighting parameters are given by:*

$$\alpha_\mathrm{i} = \frac{\mathcal{L}_\mathrm{i}}{\sqrt{\Omega_\mathrm{i}} \left(\sum_{i=1}^{N} \mathcal{L}_\mathrm{i} \sqrt{\Omega_\mathrm{i}}\right)}, \quad \forall i = 1, 2, \ldots, N. \tag{19}$$

In addition, these parameters yield the optimal step-size:

$$\mu = \frac{\sqrt{2}}{\sqrt{K} \left(\sum_{i=1}^{N} \mathcal{L}_\mathrm{i} \sqrt{\Omega_\mathrm{i}}\right)}. \tag{20}$$

Proof Minimizing the RHS of (18) w.r.t. μ yields the result of Theorem 2.1, $f^* - f(\bar{x}) \leq \frac{\mathcal{L}\sqrt{2\,\Omega}}{\sqrt{K}}$. This optimality bound is a function of $\alpha := [\alpha^1, \alpha^2, \ldots, \alpha^N]^\top$. The best optimality bound can be achieved by considering a minimization of:

$$\phi(\alpha) = \mathcal{L}^2(\alpha)\,\Omega(\alpha) = \sum_{i=1}^{N} \frac{\mathcal{L}_\mathrm{i}{}^2}{\alpha_\mathrm{i}} \sum_{i=1}^{N} \alpha_\mathrm{i}\,\Omega_\mathrm{i}\,.$$

The optimizer of $\phi(\alpha)$ needs to satisfy the following optimality condition:

$$\frac{\alpha_\mathrm{i}{}^2\,\Omega_\mathrm{i}}{\mathcal{L}_\mathrm{i}{}^2} \sum_{j=1, j\neq i}^{N} \frac{\mathcal{L}_j{}^2}{\alpha_j} = \sum_{j=1, j\neq i}^{N} \alpha_j \Omega_j, \quad \forall i = 1, 2, \ldots, N. \tag{21}$$

Now, let us rewrite the optimality bound $\frac{\Omega}{K\mu} + \frac{\mu\,\mathcal{L}^2}{2}$ in (18) as:

$$\frac{\Omega}{K\mu} + \frac{\mu\,\mathcal{L}^2}{2} = \frac{\sum_{i=1}^{N} \alpha_\mathrm{i}\,\Omega_\mathrm{i}}{K\mu} + \frac{\mu}{2} \sum_{i=1}^{N} \frac{\mathcal{L}_\mathrm{i}{}^2}{\alpha_\mathrm{i}}.$$

Minimizing the RHS of the above equality w.r.t. α_i and substituting $\mu = \frac{\sqrt{2\,\Omega}}{\mathcal{L}\sqrt{K}}$ (Theorem 2.1) in the minimizer give $\alpha_\mathrm{i} = \frac{\mathcal{L}_\mathrm{i}\sqrt{\Omega}}{\mathcal{L}\sqrt{\Omega_\mathrm{i}}}, \forall i = 1, 2, \ldots, N$. Substituting these weighting parameters into the maximum distance, $\Omega = \sum_{i=1}^{N} \alpha_\mathrm{i}\,\Omega_\mathrm{i}$, yields $\sqrt{\Omega} = \frac{\sum_{i=1}^{N} \mathcal{L}_\mathrm{i}\sqrt{\Omega_\mathrm{i}}}{\mathcal{L}}$. Suppose the weighted distance is normalized by the weighting parameters, i.e., $\Omega = 1$, then the weighted Lipschitz is given by:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_\mathrm{i}\sqrt{\Omega_\mathrm{i}}. \tag{22}$$

Using the above-weighted Lipschitz constant and the normalized maximum distance, $\Omega = 1$, yields the optimal weighting parameters (19). We can verify that the optimal α_i normalizes the maximum distance, i.e., $\Omega = 1$, generates the weighted Lipschitz constant (22) using the definition (16) and satisfies the optimality condition (21) of the optimality bound function $\phi(\alpha)$. □

Theorem 3.2 *Let f^* denotes the global optimal objective function and $\bar{x} = \mathrm{argmax}_{x=\{x^1,\ldots,x^K\}} f(x)$. The weighted MD algorithm with the optimal step-size (20) and the optimal weighting parameters (19) has the following optimality bound after K iterations:*

$$f^* - f(\bar{x}) \leq \frac{\sqrt{2}\sum_{i=1}^{N} \mathcal{L}_\mathrm{i}\sqrt{\Omega_\mathrm{i}}}{\sqrt{K}}. \tag{23}$$

Proof Substituting the optimal step-size (20) and the optimal weighting parameters (19) into (18) directly yields the result. □

The following result establishes the relative performance of the proposed weighted MD algorithm compared to the MD algorithm with unweighted distance. The proposed algorithm with weighted distance is an improvement over the algorithm with unweighted distance. Numerical experiments discussed in the next section and the supplementary material underline this promising result.

Corollary 3.2 *The optimality bound* (23) *of the proposed weighted MD algorithm is either an improvement to, or in the worst case as good as, the optimality bound* (7) *of the MD algorithm with unweighted distance:*

$$\frac{\sqrt{2}\sum_{i=1}^{N}\mathcal{L}_{\mathrm{i}}\sqrt{\Omega_{\mathrm{i}}}}{\sqrt{K}} \leq \frac{\sqrt{\sum_{i=1}^{N}\mathcal{L}_{\mathrm{i}}{}^{2}}\sqrt{2\sum_{i=1}^{N}\Omega_{\mathrm{i}}}}{\sqrt{K}}. \tag{24}$$

Proof By the Cauchy–Schwarz inequality, we have:

$$\left(\sum_{i=1}^{N}\mathcal{L}_{\mathrm{i}}\sqrt{\Omega_{\mathrm{i}}}\right)^{2} \leq \left(\sum_{i=1}^{N}\mathcal{L}_{\mathrm{i}}{}^{2}\right)\left(\sum_{i=1}^{N}\Omega_{\mathrm{i}}\right).$$

The above inequality directly yields (24). □

4 Weighted Mirror Descent Algorithm for MRF Optimization

Markov Random Fields [8] are an important class of graph-structured models in image processing and machine learning. In general, the MRF model aims to reveal hidden quantities ξ based on some observations of available input data. Various discussion about MRF modeling and MRF optimization methods in image analysis and machine learning can be found in [6,8,12,13]. In this paper, we focus on the dual of the linear programming (LP) relaxation for the MRF optimization problem. The detailed description of the MRF model and the construction of the dual problem can be found in the supplementary material provided (see also [6]). Let us consider the LP relaxation of the MRF problem:

$$\min_{\xi\in\Xi^{G}}\langle\theta,\xi\rangle. \tag{25}$$

Applying the dual decomposition technique yields the dual objective function:

$$\sum_{t\in T}\min_{\xi^{t}\in\Xi^{t}}\langle\,\theta^{t}+\lambda^{t},\xi^{t}\,\rangle\,.$$

In this setting, the sum of data cost θ^{t} must equal to the original θ (see [6] or the supplementary material):

$$\sum_{t\in T}\theta^{t}=\theta, \tag{26}$$

and the Lagrangian vector λ becomes the decision variables of the dual optimization problem:

$$\max_{\lambda\in\Lambda} \sum_{t\in T} \min_{\xi^t\in\Xi^t} \langle\, \theta^t + \lambda^t, \xi^t \,\rangle, \tag{27}$$

where $\Lambda := \left\{\sum_{t\in T} \lambda^t = \mathbf{0}\right\}$. The domain Λ is a Cartesian product of subsets $\{\Lambda_i\}_{\forall i\in I}$, where $I := \{(a,l)\}_{\forall a\in V,\forall l\in L} \bigcup \{(ab,lk)\}_{\forall ab\in E,\forall l,k\in L}$. Each subset is defined as $\Lambda_i := \left\{\sum_{t\in T} \lambda_i^t = 0\right\}, \forall i \in I$. As a result, $\Lambda = \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_{\mathcal{I}}$, where $\mathcal{I}$ is the cardinality of I. It is well known that the solution of (27) is the lower bound of the LP problem (25). By strong duality, the solution of (27) becomes the solution of the LP (25). Problem (27) is a nonsmooth convex optimization problem over the Cartesian product of convex subsets (1).

There have been several approaches for solving the nonsmooth problem (27). One approach is by Savchynskyy et al. [7] using Nesterov's smoothing technique. Their method relaxes the nonsmooth objective function by a smoothing parameter. As a result, the algorithm only computes a suboptimal solution of the dual problem and does not yield the optimal solution for the LP problem (25). In addition, this algorithm requires computations for all dual variables at every iteration, while the weighted MD requires fewer dual updates as the algorithm converges (as we will see in Remark 4.1). Schmidt et al. [14] proposed a primal-dual method for solving the LP (25); however, their paper shows that the primal-dual method is inferior to the dual decomposition technique for large-scale problem. The weighted MD algorithm is a generalization of the projected subgradient algorithm which was also proposed for solving the dual (27) by Komodakis et al. [6] and Jancsary et al. [15].

4.1 Weighted MD for the MRF Problem

Problem (27) requires an initialization of θ^t that satisfies (26). The standard initialization $\theta^t = \frac{\theta}{\mathcal{T}}$ might not give a good starting point for subgradient-typed methods. A better initialization is an initialization such that the objective function value is closer to the optimal objective value. Suppose we have a better initialization θ^{t*}, we can reduce the computational efforts for solving λ significantly. To this end, let us introduce the following optimization problem:

$$\max_{\rho\in\Delta} f(\rho) := \max_{\rho\in\Delta} \sum_{t\in T} \min_{\xi^t\in\Xi^t} \langle\, \rho^t \circ \theta\,,\, \xi^t \,\rangle, \tag{28}$$

where $\circ$ is a Hadamard product notation, $\Delta = \Delta_1 \times \Delta_2 \times \cdots \times \Delta_{\mathcal{I}}$ is the product set of simplices:

$$\Delta_i := \left\{ \rho_i : \sum_{t\in T} \rho_i^t = 1\,;\ \rho_i^t \geq 0\,,\ \forall t \in T \right\}, \quad \forall i \in I. \tag{29}$$

Problem (28) also has the same form as (1) and can be solved using the weighted MD algorithm. After obtaining the optimal initialization $\{\rho^{t*} \circ \theta\,,\ \forall t \in T\}$, where

$\rho^* = \text{argmax}_{\rho\in\Delta} f(\rho)$, we can proceed to solve for λ:

$$\max_{\lambda\in\Lambda} f(\lambda) := \max_{\lambda\in\Lambda} \sum_{t\in T} \min_{\xi^t\in\Xi^t} \langle \rho^{t*} \circ \theta + \lambda^t, \xi^t \rangle, \tag{30}$$

where $\Lambda = \Lambda \times \Lambda \times \cdots \times \Lambda_{\mathcal{I}}$ is the product set of linear subsets:

$$\Lambda_i := \left\{ \lambda_i : \sum_{t\in T} \lambda_i^t = 0 \right\}, \quad \forall i \in I. \tag{31}$$

The two problems (28) and (30) can be combined into one problem:

$$\max_{\rho\in\Delta,\lambda\in\Lambda} f(\rho,\lambda) := \max_{\rho\in\Delta,\lambda\in\Lambda} \sum_{t\in T} \min_{\xi^t\in\Xi^t} \langle \rho^t \circ \theta + \lambda^t, \xi^t \rangle. \tag{32}$$

By setting $\lambda = 0$, we have (32) $\equiv$ (28). Similarly, if we set $\rho^{t*} = \text{argmax}_{\rho\in\Delta} f(\rho)$, then we have (32) $\equiv$ (30). The weighted MD algorithm for solving the MRF problem is described in Algorithm 1. As we will see later (equation (40)), exact and optimal step-size τ can be computed while the exact η is not available. A heuristic based on the difference between the current objective value and the optimal solution will be used to approximate η. The smaller this difference is, the less error accumulates in approximating λ. Therefore, the solution to problem (28) yields a starting point for λ such that its objective value is closer to the optimal solution compared to an objective value corresponding to a random starting point. We clarify the various aspects of the vector ρ (similar for λ):

Algorithm 1: Weighted Mirror Descent for the MRF Problem

Step 1: Choose two nonegative numbers K_1, K_2;
Step 2: Initialize $\rho^1 = \frac{1}{\mathcal{I}}.\mathbf{1}$ and $\lambda^1 = \mathbf{0}$;
Step 3:
for $k = 1, 2, \ldots, K_1 - 1$ **do**

$$\rho^{k+1} = \underset{\rho\in\Delta}{\text{argmax}} \langle f'_{\rho^k}, \rho \rangle - \frac{1}{\tau} \mathrm{D}_\Delta(\rho, \rho^k). \tag{33a}$$

Step 4: Set $\bar{\rho} = \underset{\rho}{\text{argmax}} \left\{ f(\rho, \lambda^1) \,\middle|\, \rho = \rho^1, \rho^2, \ldots, \rho^{K_1} \right\}$;
Step 5:
for $k = 1, 2, \ldots, K_2 - 1$ **do**

$$\lambda^{k+1} = \underset{\lambda\in\Lambda}{\text{argmax}} \langle f'_{\lambda^k}, \lambda \rangle - \frac{1}{\eta} \mathrm{D}_\Lambda(\lambda, \lambda^k). \tag{33b}$$

Step 6: Set $\bar{\lambda} = \underset{\lambda}{\text{argmax}} \left\{ f(\bar{\rho}, \lambda) \,\middle|\, \lambda = \lambda^1, \lambda^2, \ldots, \lambda^{K_2} \right\}$;

- $\rho \in \Delta$ denotes a full vector corresponding to all subgraphs of the set T.
- With superscript t, ρ^t denotes a vector corresponding to subgraph $t \in T$.
- With subscript i, ρ_i denotes a collection of scalars ρ_i^t across all subgraphs that cover the index i, and $\rho_i \in \Delta_i$.
- With numeric superscripts, $\rho^1, \rho^2, .., \rho^K$, or ρ^k, ρ_i^k denote the corresponding iterate of the vector.
- When superscripts t and k are used together, we separate them by a comma: $\rho^{t,k}$ is a vector, or $\rho_i^{t,k}$ is a scalar.

The two weighted distances D_Δ and D_Λ yield the corresponding subset projections for (33):

$$\forall i \in I: \quad \rho_i^{k+1} = \underset{\rho_i \in \Delta_i}{\operatorname{argmax}} \left\langle f'_{\rho_i^k}, \rho_i \right\rangle - \frac{\alpha_{\Delta_\mathrm{i}}}{\tau} \mathrm{D}_{\Delta_\mathrm{i}}(\rho_i, \rho_i^k). \tag{34a}$$

$$\forall i \in I: \quad \lambda_i^{k+1} = \underset{\lambda_i \in \Lambda_i}{\operatorname{argmax}} \left\langle f'_{\lambda_i^k}, \lambda_i \right\rangle - \frac{\alpha_{\Lambda_\mathrm{i}}}{\eta} \mathrm{D}_{\Lambda_\mathrm{i}}(\lambda_i, \lambda_i^k). \tag{34b}$$

To this end, we choose the log-entropy distance function for each subset Δ_i and the Euclidean distance function for each subset Λ_i. Let us consider:

- For each Δ_i: Let $\psi_\Delta^i(\rho_i) = \sum_{t \in T} \rho_i^t \log \rho_i^t$, if $\rho_i \in \Delta_i$; $else, +\infty$. Then, ψ_Δ^i is 1-strongly convex [4, Proposition5.1] w.r.t. $\|.\|_1$. The dual norm of $\|.\|_1$ is $\|.\|_\infty$ [11].
- For each Λ_i: Let $\psi_\Lambda^i(\lambda_i) = \frac{1}{2} \sum_{t \in T} (\lambda_i^t)^2$, if $\lambda_i \in \Lambda_i$; $else, +\infty$. Then, ψ_Λ^i is 1-strongly convex w.r.t. $\|.\|_2$. The dual norm of $\|.\|_2$ is itself.

By using the Bregman distance, we can obtain the log-entropy distance function and the Euclidean distance function for the corresponding subset. As a result, each iteration of the recurrences (34) can be solved in a closed form:

$$\forall i \in I: \quad \rho_i^{t,k+1} = \frac{\rho_i^{t,k} \times \exp\left(\frac{\tau}{\alpha_{\Delta_\mathrm{i}}} \times f'_{\rho_i^{t,k}}\right)}{\sum_{t \in T} \left(\rho_i^{t,k} \times \exp\left(\frac{\tau}{\alpha_{\Delta_\mathrm{i}}} \times f'_{\rho_i^{t,k}}\right)\right)}. \tag{35a}$$

$$\forall i \in I: \quad \lambda_i^{t,k+1} = \frac{\eta}{\alpha_{\Lambda_\mathrm{i}}} \left(f'_{\lambda_i^{t,k}} - \frac{\sum_{t \in T} f'_{\lambda_i^{t,k}}}{\mathcal{T}} \right). \tag{35b}$$

We note that MD algorithm with unweighted distance also uses the above recurrences with the constant choice $\alpha_{\Delta_\mathrm{i}} = \alpha_{\Lambda_\mathrm{i}} = 1, \forall i \in I$. Using the definitions of optimal step-size (20) and weighting pararmeters (19), the two subset-dependent step-sizes $\frac{\tau}{\alpha_{\Delta_\mathrm{i}}}$ and $\frac{\eta}{\alpha_{\Lambda_\mathrm{i}}}$ can be written as:

$$\frac{\tau}{\alpha_{\Delta_\mathrm{i}}} = \frac{\sqrt{2\,\Omega_{\Delta_\mathrm{i}}}}{\mathcal{L}_{\Delta_\mathrm{i}}\sqrt{k}} \quad \text{and} \quad \frac{\eta}{\alpha_{\Lambda_\mathrm{i}}} = \frac{\sqrt{2\,\Omega_{\Lambda_\mathrm{i}}}}{\mathcal{L}_{\Lambda_\mathrm{i}}\sqrt{k}}. \tag{36}$$

The above subset-dependent step-sizes improve the performance of the weighted MD because they use optimal values of α_{Δ_i} and α_{Λ_i} instead of the constant 1. It thus remains to show how to compute the subgradients f'_ρ and f'_λ at any feasible $\rho \in \Delta$ and $\lambda \in \Lambda$.

Lemma 4.1 *Let* $\bar{\xi}^t = \text{argmin}_{\xi^t \in \Xi^t} \langle \rho^t \circ \theta + \lambda^t, \xi^t \rangle$ *be the optimal solution for the MRF subproblem of the corresponding subgraph* $t \in T$*. Then, the subgradients of* $f(\rho, \lambda)$ *w.r.t. the corresponding decision vector are given by:*

$$f'_{\rho^t} = \theta \circ \bar{\xi}^t \quad \text{and} \quad f'_{\lambda^t} = \bar{\xi}^t.$$

Proof Let x, y be arbitrary vectors such that $x \in \Delta$ and $y \in \Lambda$. By definition, $\bar{\xi}^t$ is not necessarily optimal for $\min_{\xi^t \in \Xi^t} \langle x^t \circ \theta + y^t, \xi^t \rangle$, i.e.,

$$\forall t \in T: \quad \min_{\xi^t \in \Xi^t} \langle x^t \circ \theta + y^t, \xi^t \rangle \leq \langle x^t \circ \theta + y^t, \bar{\xi}^t \rangle .$$

In addition,

$$\begin{aligned} f(x, y) &= \sum_{t \in T} \min_{\xi^t \in \Xi^t} \langle x^t \circ \theta + y^t, \xi^t \rangle \leq \sum_{t \in T} \langle x^t \circ \theta + y^t, \bar{\xi}^t \rangle \\ &= F(\rho, \lambda) + \langle \theta \circ \bar{\xi}, x - \rho \rangle + \langle \bar{\xi}, y - \lambda \rangle. \end{aligned}$$

□

Remark 4.1 The above choices of subgradient rely on the exact solution $\bar{\xi}^t \in \Xi^{\mathcal{I}}$ for each subgraph t (that can be computed very efficiently by a dynamic programming algorithm, e.g., max-product belief propagation or graph cut). Using these subgradients, we can verify that updates (35) are only needed at *disagreement nodes*.[2] As a result, we can utilize this property to define a stopping criterion by counting the number of disagreement nodes. Let L_k be the number of disagreement nodes at iteration k. Essentially, as $L_k \to 0$, the algorithm converges to a stationary point, i.e., the optimal solution.

By using the above subgradients and the fact that $\bar{\xi}^t_i \in [0, 1]$, we can derive the local Lipschitz constants corresponding to their subsets, $\forall i \in I$:

$$\mathcal{L}_{\Delta_i} = \sup_{\rho_i \in \Delta_i} \| f'_{\rho_i} \|_\infty = |\theta_i| \quad \text{and} \quad \mathcal{L}_{\Lambda_i} = \sup_{\lambda_i \in \Lambda_i} \| f'_{\lambda_i} \|_2 = \sqrt{\mathcal{T}}. \tag{37}$$

To specify the maximum subset distances, we need to find an upper bound for the distance between any feasible point to starting points ρ_i^1 and λ_i^1.

[2] A node $a \in V$ is a *disagreement node* if all subgraphs do not assign the same label to a, i.e., for any two subgraphs $t_1, t_2 \in T$, there exists $l \in L$ such that $\bar{\xi}^{t_1}_{a,l} \neq \bar{\xi}^{t_2}_{a,l}$.

Lemma 4.2 *Let all elements of starting point $\rho_i^{t,1} = \frac{1}{\mathcal{T}}$, and the upper bound of the distance between any feasible vector and ρ_i^1 is given by:*

$$\Omega_{\Delta_i} = \log \mathcal{T}. \tag{38}$$

Proof Using the Bregman distance (6) with log-entropy function $\psi_\Delta^i(\rho_i) = \sum_{t\in T} \rho_i^t \log \rho_i^t$ for every subset $\Delta_i, i \in \mathcal{I}$, we have:

$\mathrm{D}_{\Delta_i}(\rho_i, \rho_i^1) = \sum_{t\in T} \rho_i^t \log \rho_i^t + \left(\sum_{t\in T} \rho_i^t\right) \log \mathcal{T} \le \left(\sum_{t\in T} \rho_i^t\right) \log \mathcal{T} \le \log \mathcal{T}$. The last two inequalities follow from the facts that $0 \le \rho_i^t \le 1$; therefore, $\log \rho_i^t \le 0$, and $\sum_{t\in T} \rho_i^t = 1$. □

Similar to the above, the Bregman distance with $\psi_\Lambda^i(\lambda_i) = \frac{1}{2}\sum_{t\in T}(\lambda_i^t)^2$ yields the Euclidean distance corresponding to subset Λ_i; thus, the quantity Ω_{Λ_i} is given by (with $\lambda_i^1 = \mathbf{0}$) $\Omega_{\Lambda_i} = \max_{\lambda_i\in\Lambda_i} \frac{1}{2}\|\lambda_i - \lambda_i^1\|_2^2 = \max_{\lambda_i\in\Lambda_i} \frac{1}{2}\|\lambda_i\|_2^2$. The subset Λ_i defined in (31) does not allow exact computation for Ω_{Λ_i}. For example, assume the index $i \in I$ is covered by two subgraphs $t_1, t_2 \in T$, then

$$2\,\Omega_{\Lambda_i} = \max_{\lambda_i^{t_1}+\lambda_i^{t_2}=0} \|\lambda_i\|_2^2 = \max_{\lambda_i^{t_1}+\lambda_i^{t_2}=0} (\lambda_i^{t_1})^2 + (\lambda_i^{t_2})^2.$$

The quantity $2\,\Omega_{\Lambda_i}$ can be infinitely large. Thus, the step-size $\frac{\eta}{\alpha_{\Lambda_i}}$ also becomes infinitely large. In this problem, we assume subset Λ_i to be bounded and nonempty. Therefore, we estimate Ω_{Λ_i} by a quantity that is proportional to the distance between the solution λ_i^* and the starting point $\lambda_i^1 = \mathbf{0}$. Given the primal problem (25) and dual problem (32), we use the approximate duality gap (since the primal solutions cannot always be computed exactly using the dual solutions) as a heuristic estimation of the distance between the current iterate and the optimal solution.

In order to estimate the duality gap at iteration k, we need to compute (approximately) the primal value $P(\xi^k) = \langle\, \theta, \xi^k \,\rangle$. Several approaches to estimate the primal variables are discussed in [6]. We employ the ergodic sequence of dual subgradients f'_{λ^k} to estimate the primal variables. Ergodic convergence analysis [16] has been used by many authors to bridge the primal-dual gap in convex optimization. In the approach, primal variables ξ^k are estimated by considering the weighted average of the dual subgradients over all iterations:

$$\xi^K = \frac{\sum_{k=1}^K \sum_{t\in T} f'_{\lambda^{t,k}}}{K} = \frac{\sum_{k=1}^K \sum_{t\in T} \bar{\xi}^{t,k}}{K}.$$

The approximate duality gap is given by $|P(\xi^K) - f(\bar{\rho}, \lambda^K)|$, which can be used as a heuristic to estimate Ω_{Λ_i} at iteration k:

$$\Omega_{\Lambda_i} = \frac{|P(\xi^k) - f(\bar{\rho}, \lambda^k)|}{2L_k}, \tag{39}$$

where L_k is the number of disagreement nodes (see Remark 4.1). Substituting local Lipschitz constants (37) and subset distances (38), (39) into the subset-dependent step-sizes (36) yields:

$$\frac{\tau}{\alpha_{\Delta_i}} = \frac{\sqrt{2\log(\mathcal{T})}}{|\theta_i|\sqrt{k}} \quad \text{and} \quad \frac{\eta}{\alpha_{\Lambda_i}} = \sqrt{\frac{|P(\xi^k) - f(\bar{\rho}, \lambda^k)|}{L_k\, \mathcal{T}\, k}}. \tag{40}$$

Relating the step-size $\frac{\eta}{\alpha_{\Lambda_i}}$ to the duality gap allows the algorithm to admit large step-sizes when the duality gap is large (far from the optimum). As the duality gap reduces, so does the step-size. This choice of step-size is consistent with the diminishing step-size approach that guarantees convergence for subgradient methods [17].

4.2 Numerical Experiments

Experimental results are discussed in the supplementary material provided and published online along with this paper.

5 Conclusions

An efficient algorithm is presented for solving a large-scale nonsmooth convex problem. The method is based on the Mirror Descent algorithm employing a suitable weighted distance function. By assessing the optimality bound of the proposed algorithm, we are able to compute the optimal subset-dependent step-sizes. This yields a convergence rate that is not worse than the MD algorithm with unweighted distance. The experimental results for MRF optimization problems confirm the improved performance.

Acknowledgments We acknowledge a partial support of the EPSRC award EP/I014640/1 for the author Duy V.N. Luong. The work of the second author was partially supported by the FP7 Marie Curie Career Integration Grant (PCIG11-GA-2012-321698 SOC-MP-ES) and the EPSRC Grant EP/K040723/1.

Compliance with Ethical Standards

Conflict of interest All the authors do not have any Conflict of interest.

References

1. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley, Chichester (1983)
2. Juditsky, A., Nemirovski, A.: First order methods for nonsmooth convex large-scale optimization, i: General purpose methods, chap. 5. In: Sra, S., Nowozin, S., Wright, S.J. (eds.) Optimization for Machine Learning. The MIT Press, Cambridge (2012)
3. Ben-tal, A., Margalit, T., Nemirovski, A.: The ordered subsets mirror descent optimization method with applications to tomography. SIAM J. Optim. **12**, 2001 (2001)
4. Beck, A., Teboulle, M.: Mirror descent and nonlinear projected subgradient methods for convex optimization. Oper. Res. Lett. **31**(3), 167–175 (2003)

5. Kiwiel, K.C.: Proximal minimization methods with generalized bregman functions. SIAM J. Control Optim. **35**(4), 1142–1168 (1997)
6. Komodakis, N., Paragios, N., Tziritas, G.: MRF energy minimization and beyond via dual decomposition. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 531–552 (2011)
7. Savchynskyy, B., Schmidt, S., Kappes, J., Schnorr, C.: A study of nesterov's scheme for lagrangian decomposition and map labeling. In: Computer Vision and Pattern Recognition, pp. 1817–1823 (2011)
8. Li, S.Z.: Markov Random Field Modelling in Image Analysis. Advances in Computer Vision and Pattern Recognition. Springer, London (2009)
9. Censor, Y., Zenios, S.A.: Proximal minimization algorithm with d-functions. J. Optim. Theory Appl. **73**, 451–464 (1992)
10. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM J. Optim. **3**, 538–543 (1993)
11. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
12. Kumar, M.P., Kolmogorov, V., Torr, P.H.S.: An analysis of convex relaxations for MAP estimation of discrete MRFs. J. Mach. Learn. Res. **10**, 71–106 (2009)
13. Sontag, D., Globerson, A., Jaakkola, T.: Introduction to dual decomposition for inference. In: Sra, S., Nowozin, S., Wright, S.J. (eds.) Optimization for Machine Learning. MIT Press, Cambridge (2011)
14. Schmidt, S., Savchynskyy, B., Kappes, J.H., Schnorr, C.: Evaluation of a first-order primal-dual algorithm for mrf energy minimization. In: EMMCVPR (2011)
15. Jancsary, J., Matz, G.: Convergent decomposition solvers for tree-reweighted free energies. J. Mach. Learn. Res. (15) 388–398 (2011)
16. Larsson, T., Patriksson, M., Stromberg, A.: Ergodic primal convergence in dual subgradient schemes for convex programming. Math. Program. **86**, 283–312 (1999)
17. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont (1999)